Stone, Glenn; Clifford, David; Gustafsson, Ove Johan Ragnar; McColl, Shaun Reuss; Hoffmann, Peter
Visualisation in imaging mass spectrometry using the minimum noise fraction transform, BMC Research Notes, 2012; 5:419

Stone *et al. BMC Research Notes* 2012, **5**:419
http://www.biomedcentral.com/1756-0500/5/419

BMC Research Notes

TECHNICAL NOTE Open Access

Visualisation in imaging mass spectrometry using the minimum noise fraction transform

Glenn Stone[1*], David Clifford[2], Johan OR Gustafsson[3], Shaun R McColl[3] and Peter Hoffmann[3]

Abstract

Background: Imaging Mass Spectrometry (IMS) provides a means to measure the spatial distribution of biochemical features on the surface of a sectioned tissue sample. IMS datasets are typically huge and visualisation and subsequent analysis can be challenging. Principal component analysis (PCA) is one popular data reduction technique that has been used and we propose another; the minimum noise fraction (MNF) transform which is popular in remote sensing.

Findings: The MNF transform is able to extract spatially coherent information from IMS data. The MNF transform is implemented through an R-package which is available together with example data from http://staff.scm.uws.edu.au/~glenn/#Software.

Conclusions: In our example, the MNF transform was able to find additional images of interest. The extracted information forms a useful basis for subsequent analyses.

Keywords: Dimension reduction, MALDI imaging mass spectrometry, Image processing

Background

Imaging Mass Spectrometry (IMS) provides a means to measure the spatial distribution of drug metabolite, lipid, peptide and protein features on the surface of a sectioned tissue sample (see [1] and references therein). Typically, IMS methods utilise freshly frozen sections of tissue mounted onto conductive slides. These are coated with matrix followed by MALDI-ToF/ToF spectra acquisition at anywhere from hundreds to thousands of positions across a tissue, the spatial locations of which are annotated. For example, a section of coronal murine midbrain can generate more than $\sim$ 2000 spectra. Data acquisition at 0.1 GS/s over an m/z range 1000-26000 yields individual mass spectra with more than 11,000 plotted points. The resulting data set is enormous and thus difficult to process, visualise and analyse effectively.

The data can be thought of in two ways, firstly a set of mass spectra acquired at a spatial array of spots, and secondly as a *stack* of ion intensity maps, each map being akin to a low resolution image. Software such as Biomap and flexImaging (Bruker Daltonics) view IMS data as ion intensity maps and include features such as data normalisation and noise spectra exclusion (see Figure 1). However the choice of ion intensity maps to view is largely user driven and images are noisy. Further data analysis using external software packages is possible, for example, (ClinProTools for principal component analysis (PCA), hierarchical clustering (HC) of spectra, or spectral model generation [2-5]. Other analysis techniques used on IMS data include kriging of ion intensity maps [6] and supervised classification methods, for example, random forests [7].

Current methods typically use spectral features, not spatial information, to guide analysis. Hence the predominance of PCA and HC type approaches. We propose the use of the minimum noise fraction (MNF) transform [8] to, firstly, determine the most interesting spatial representations of IMS data, and secondly, form the basis of data reduction for subsequent analysis. The MNF transform has previously been used on hyper-spectral images of tissue samples [9] but this is the first use of such a technique on IMS data.

Findings

Principal Components Analysis

Principal Components Analysis (PCA) treats the IMS data as a collection of spectra. Therefore, in PCA, the spatial

*Correspondence: g.stone@uws.edu.au
[1] School of Computing, Engineering and Mathematics, University of Western Sydney, Sydney, New South Wales, Australia
Full list of author information is available at the end of the article

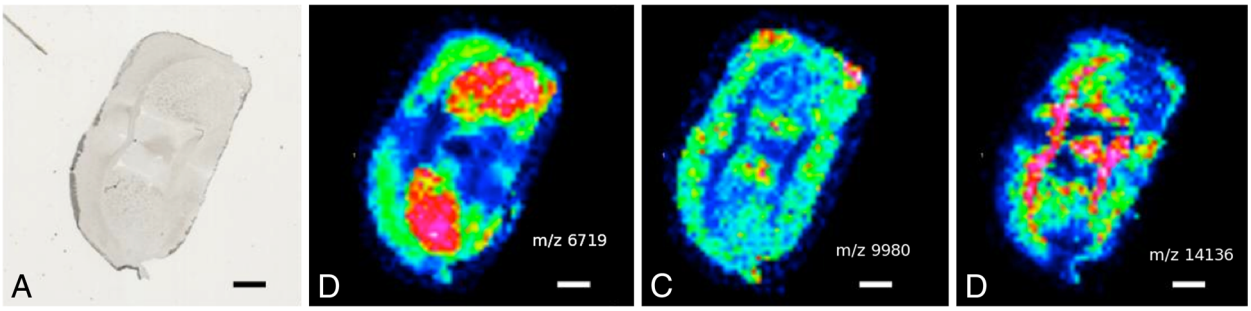

Figure 1 Image (A) of coronal murine midbrain section, and ion intensity maps at m/z of 6719 (B), 9980 (C) and 14136 (D). m/z are approximate, from flexImaging V2.1. Scale bar is 1 mm.

structure of the spots is not relevant and so the data can be represented as a matrix $\mathcal{Z} = \{Z_{ik}\}$ where $i = 1, \ldots, n$ ranges over the spots on the tissue and $k = 1, \ldots, p$ ranges over the mass charge ratios in the mass spectrum. Let $Z = \{Z_k\}$ be a typical mass spectrum. PCA seeks linear combinations of intensities over the mass charge ratios that maximizes variance. That is, the first principal component is defined by a vector $a = \{a_k\}$ with a chosen so that $\mathrm{Var}(a^t Z)$ is maximised. Second and subsequent principal components maximize variance subject to being uncorrelated with all previous principal components.

If Σ_Z is the covariance matrix of the mass spectra, that is, the (k_1, k_2) entry is the covariance of the ion intensity measured at the k_1-th m/z ratio and the ion intensity measured at the k_2-th m/z ratio, then the first principal component maximises $a^t \Sigma_Z a$ subject to a suitable scale constraint such as $a^t a = 1$. Generally, Σ_Z is unknown so is estimated using the sample covariance matrix S_Z given by

$$(S_Z)_{k_1 k_2} = \frac{1}{n-1} \sum_i \left(Z_{ik_1} - \bar{Z}_{\cdot k_1}\right)\left(Z_{ik_2} - \bar{Z}_{\cdot k_2}\right)$$

where

$$\bar{Z}_{\cdot k} = \frac{1}{n} \sum_i Z_{ik}$$

It should be noted that the mass spectra are unlikely to form a set of independent observations since spatially close spectra will likely be correlated.

The MNF transform

PCA makes no use of the spatial structure of the observed mass spectra. The Minimum Noise Fraction transform uses a simple model to allow the spatial structure to influence the analysis. Here we modify the notation to emphasise the spatial aspect; let $Z(x)$ be the mass spectrum at spatial location x. In our case, x will be a spot on the tissue section indexed by a horizontal and a vertical coordinate. A possible model for $Z(x)$ is

$$Z(x) = M(x) + N(x)$$

where $M(x)$ represents the *signal* at x and $N(x)$ is the *noise* at x.

This is to be interpreted as "the mass spectrum at spot x is composed of a spatial signal mass spectrum plus a noise mass spectrum". We assume the signal and noise components to be independent, and the noise component to have low spatial covariance. The signal component would likely have high spatial covariance. Both components would still have a covariance between intensities at differing mass charge ratios, represented by covariance matrices Σ_M and Σ_N.

The MNF transform seeks linear combinations of intensities over the mass charge ratios that maximizes *signal to noise ratio* (SNR). That is, the first MNF band is defined by a vector $a = \{a_k\}$ with a chosen so that $\mathrm{SNR} = \mathrm{Var}(a^t M)/\mathrm{Var}(a^t N)$ is maximised. Replacing the variances by expressions in terms of the covariance matrices we see that;

$$\mathrm{SNR} = \frac{a^t \Sigma_M a}{a^t \Sigma_N a}$$

In PCA we need an estimate for Σ_Z, whereas for the MNF we need estimates of Σ_M and Σ_N. These are not as straight-forward to obtain as in PCA, since the signal M and noise N components are not directly observed. However, by noting that (by independence) $\Sigma_Z = \Sigma_M + \Sigma_N$ we see that the SNR is maximised when the following ratio is maximised,

$$\frac{a^t \Sigma_Z a}{a^t \Sigma_N a}$$

Thus only an estimate for Σ_N is required. In reality, only an estimate of Σ_N *or* Σ_M is required, and we find it easiest to estimate the former.

Green et al. [8] propose a *shift difference* method to estimate Σ_N and Berman et al. [9] propose using the covariance of residuals from a local quadratic fit. In the latter case, a quadratic function is fit to a 3×3 neighbourhood of each spot for each mass charge ratio, and a residual computed at the spot. This produces a set of pseudo-residual data and the sample covariance of this used as the estimate. These original applications of the MNF transform are based on hyperspectral images where the spots are very close together. Here the spots from which MS spectra are collected are somewhat separated. For this reason we have used a simpler local linear fit, based on the ideas of [10], which is similar to using a symmetrical set of shift differences. Using the simpler approach places less reliance on spots that are further apart. Although there is scope to investigate other approaches, preliminary work shows little difference when a quadratic signal fit is used in this case.

Each spot (except edge spots) has two horizontal and two vertical neighbours. Averaging these four values gives the prediction of a local linear fit at the central spot, from which a pseudo-residual can be derived. Since the spots are on a regular grid, this corresponds to the residual from a local linear fit to the four neighbouring spots. This procedure produces a set of pseudo-residuals (one for each spot at each mass charge ratio, subject to simple modification at edge spots) from which the sample noise covariance matrix S_N can be formed. We use this as the estimate of Σ_N.

Implementation

PCA corresponds to the maximisation of $a^t S_Z a$ subject to a scale constraint such as $a^t a = 1$. Lagrange multipliers can be used to show that at the maximum, a is the eigenvector of S_Z corresponding to the largest eigenvalue. Subsequent principal components are defined by eigenvectors corresponding to subsequent eigenvalues. A similar argument shows that the a which maximises the ratio

$$\frac{a^t S_Z a}{a^t S_N a}$$

satisfies

$$S_Z a = \lambda S_N a$$

This is a *generalised eigenproblem* (see [11] for example).

For both PCA and MNF the uses of all mass charge ratios would produce sample covariance matrices that are extremely large, so firstly some pre-filtering is used. In PCA, this is often a peak identification method, or selection by taking all the mass charge ratios for which the intensity exceeds some threshold (in some or all spots). For the MNF transform, we use only those mass charge ratios whose SNR values exceed a threshold. This SNR corresponds to the ratio of diagonal entries in S_Z and S_N. The threshold is chosen so that the matrices are of a manageable size.

Our implementation uses the LAPACK [12] routines for generalised eigenproblem interfaced to the R system for statistical programming [13]. All aspects of this process are automated in our code. The only manual interventions required have to do with pre-filtering of the signals and the choice of the number of bands for subsequent analysis. Both of these manual interventions are required by PCA also.

Results

We demonstrate the method using a section of 10 μm coronal murine midbrain. The section was desiccated for 30 minutes followed by washing in 70% and 100% EtOH for 30 seconds each, and subsequently desiccated until use. 20 mg/mL of 2,5-dihydroxybenzoic acid in 50% MeOH and 0.2% TFA matrix was deposited using 3 phases on an ImagePrep station. Mass spectrometry analysis

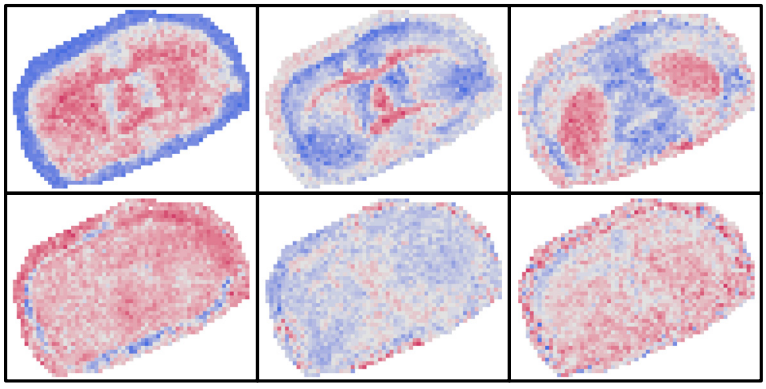

Figure 2 The first six principal component images of a coronal murine midbrain section.

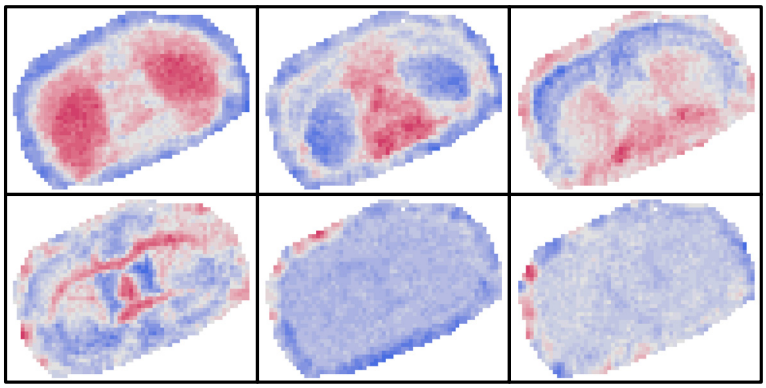

Figure 3 The first six MNF transformed images of a coronal murine midbrain section.

was carried out on an UltraFlex III MALDI-TOF/TOF machine in linear positive ion mode. ClinProT calibrants (1:20 dilution with matrix) were used to calibrate the instrument. Data acquisition used flexControl V3.3, with 300 shots taken at each spot and summed. Mass spectra were acquired in m/z range 1000–26000 at a rate of 0.1 GS/s. Figure 1 shows the image of the section (A) and flexImaging (Bruker Daltonics) generated ion intensity maps for three m/z ratios (B–D).

The processed data consists of intensities at 11280 mass charge ratios, repeated across a grid of 2012 spots over the tissue slice. The data were first logged and then background corrected by using a 5-knot robust spline fit to estimate baseline. Pre-filtering of mass charge ratios was carried by thresholding intensities (in the case of PCA) or SNRs (in the case of MNF) so that 650 were retained. PCA and MNF transforms were computed. This means that PCA operated on the 650 mass charge ratios with the highest intensity, whereas MNF used the 650 mass charge ratios with the highest estimated signal to noise. The choice of 650 data points stems from trial and error and a pragmatic desire to use manageable covariance matrices.

Figure 2 shows the first six principal component images of the data, there appears to be three images with significant spatial structure, and the remaining three appear to be noise.

Figure 3 shows the first six MNF bands. There are four images with clear spatial structure, so the MNF transform has been able to extract further information.

Subsequent Analysis

Deininger et al. [3] show the use of the principal components in hierarchical clustering, and this can also be done with the MNF bands. Hierarchical clustering is useful for identifying regions of the tissue with relatively homogenous properties. Using the PCA or MNF bands significantly reduces the computational complexity of clustering without overly reducing its usefulness.

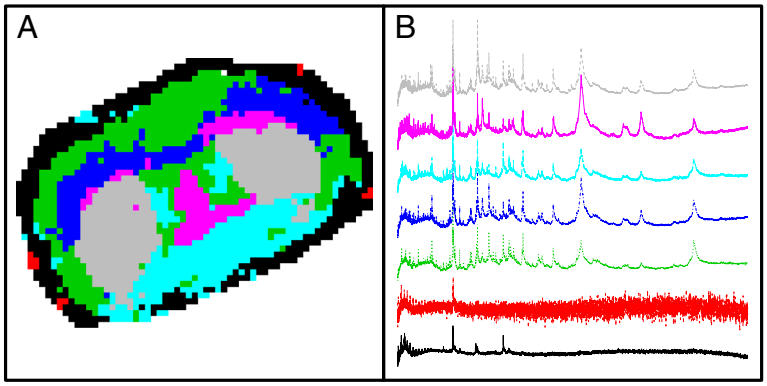

Figure 4 The results of clustering using the first four MNF bands. (A) the clustering of spots (spots in the same cluster have the same colour), **(B)** the average (background corrected) mass spectrum for each cluster.

Taking the first four MNF band images, and applying hierarchical clustering we can determine seven clusters. Figure 4 shows the results. Panel A shows the spots coloured according to cluster and panel B shows the average mass spectrum for the spots in each cluster.

As with PCA, the choice of the number of MNF bands to use in subsequent analysis (such as hierarchical clustering) is somewhat ad-hoc and depends on the form of such analysis. For clustering and classification there are many methods for choosing the number of features but we regard this as a topic for further research.

In this instance the number of components chosen (six) was primarily chosen for convenience and subjective reasons. The 5th and 6th PCA plots still show some faint internal structure, whereas subsequent ones do not (not shown). So we use 6 components for both PCA and MNF for consistency.

More generally the number of PCA components *can* be chosen using percent total variation explained arguments. In this approach, the sum of the eigenvalues for the chosen components divided by the sum of *all* the eigenvalues, converted to a percentage, is considered. A threshold percent (eg. 80 or 90%) is then chosen and the number of principal components fixed at that which first exceeds the threshold. It is not so easy to apply this technique for MNF as the eigenvalues represent signal to noise ratios and as such are not additive. However, since they are signal to noise ratios, they are scale-free and can be subject to thresholds themselves ie. take all components with eigenvalue (signal-to-noise ratio) greater than a threshold. Examples of such a threshold might be one, ie. signal and noise are approximately equal.

Conclusion

We have shown that the minimum noise fraction transform is a potent addition to the suite of analysis tools available for the analysis of Imaging Mass Spectrometry data. Like PCA, we have further demonstrated that the MNF bands generated can be used as summaries of the mass spectra to analyse the spatial characteristics of a tissue slice. We regard the MNF transform as providing a useful alternative to PCA in Imaging Mass Spectrometry. Its defining feature is that is uses estimates of spatial signal to noise ratio to sequentially define bands whereas PCA uses only total variation (signal plus noise).

Both PCA and MNF are computationally efficient when compared to the data acquisiton and preprocessing steps involved. In our implementation, all code was written in R and C and is therefore platform independent. However, the flexImaging provided data in a proprietary format that required the use of a Windows only proprietary tool (CompassXport). We have successfully used emulation software on Linux and Mac OS X based systems to run this tool.

Availability and requirements

Project Name: Computing Minimum Noise Fraction Transforms of Imaging Mass Spectrometry Data;
Project Home: http://staff.scm.uws.edu.au/~glenn/#Software;
Operating Systems: MNF code is in R and C and is compatible with Windows, Mac, and Linux;
Programming Language: R, http://cran.r-project.org and C;
Other Requirements: caMassClass,[14]; Data Acquistion and conversion software (flexImaging/CompassXport);
License GPL-2;
Restrictions to use by non-academics: none;

Availability of supporting data

The software and supporting data are available for download from the project home at http://staff.scm.uws.edu.au/~glenn/#Software.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

GS and DC concieved the statistical approach. DC implemented the analysis. GS drafted the manuscript. JORG, SRM and PH developed the protocols, and obtained, prepared and processed the samples. All authors read and approved the final manuscript.

Funding

GS was employed by CSIRO when much of this work was carried out. The Adelaide Proteomics Centre was partially funded by Bioplatforms Australia and an NHMRC equipment grant.

Acknowledgements

The authors would like to thank Mike Buckley for providing critical feedback of an early draft of this manuscript.

Author details

[1]School of Computing, Engineering and Mathematics, University of Western Sydney, Sydney, New South Wales, Australia. [2]Division of Mathematics, Informatics and Statistics, CSIRO, Brisbane, Queensland, Australia. [3]Adelaide Proteomics Centre, School of Molecular and Biomedical Science, The University of Adelaide, Adelaide, South Australia, Australia.

Received: 2 September 2011 Accepted: 25 July 2012
Published: 7 August 2012

References

1. Gustafsson J, Oehler M, Ruszkiewicz A, McColl S, Hoffmann P: **MALDI imaging mass spectrometry (MALDI-IMS) — Application of spatial proteomics for ovarian cancer classification and diagnosis.** *Int J Mol Sci* 2011, **12:**773–794.
2. Beisinger M, Paepegaey P, McIntyre N, Harbottle R, Petersen N: **Principal component analysis of TOF-SIMS images of organic monolayers.** *Anal Chem* 2002, **74:**5711–5716.
3. Deininger S, Ebert M, Futterer A, Gerhard M, Rocken C: **MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers.** *J Proteome Res* 2008, **7:**5230–5236.
4. Franck J, Arafah K, Elayed M, Bonnel D, Vergara D, Jacquet A, Vinatier D, Wisztorski M, Day R, Fournier I, Salzet M: **MALDI imaging mass spectrometry.** *Mol Cell Proteomics* 2009, **8.9:**2023–2033.
5. Smentkowski V, Ostrowski S, Kollmer F, Schnieders A, Keenan M, Ohlhausend J, Kotulad P: **Multivariate statistical analysis of**

non-mass-selected ToF-SIMS data. *Surf Interface Anal* 2005, **40:**1176–1182.
6. Milillo T, Gardella JJr: **Spatial statistics and interpolation methods for TOF SIMS imaging.** *Appl Surf Sci* 2006, **252:**6883–6890.
7. Hanselmann M, Kothe U, Kirchner M, Renard B, Amstalden E, Glunde K, Heeren R, Hamprecht F: **Toward digital staining using imaging mass spectrometry and random forests.** *J Proteome Res* 2009, **8**(7):3558–3567.
8. Green A, Berman M, Switzer P, Craig M: **A transformation for ordering multispectral data in terms of image quality with implications for noise removal.** *IEEE Trans Geoscience Remote Sensing* 1988, **26:**65–74.
9. Berman M, Phatak A, Lagerstrom R, Wood B: **ICE: a new method for the multivariate curve resolution of hyperspectral images.** *J Chemom* 2009, **23:**101–116.
10. Buckley M, Eagleson G: **A graphical method for estimating the residual variance in nonparametric regression.** *Biometrika* 1989, **76:**203–210.
11. Golub G, Van Loan, C: *Matrix Computations*. 3rd edition. Baltimore, Maryland: The Johns Hopkins University Press; 1996.
12. Anderson E, Bai Z, Bischof C, Blackford S, Demmel J, Dongarra J, Du Croz, J, Greenbaum A, Hammarling S, McKenney A, Sorensen D: *LAPACK Users' Guide*. 3rd edition. Philadelphia, PA: Society for Industrial and Applied Mathematics; 1999.
13. R Development Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2012. http://www.R-project.org/. [ISBN 3-900051-07-0].
14. Tuszynski J: **caMassClass: Processing & Classification of Protein Mass Spectra (SELDI) Data** 2010. http://CRAN.R-project.org/package=caMassClass. [R package version 1.9].

doi:10.1186/1756-0500-5-419
Cite this article as: Stone *et al.*: Visualisation in imaging mass spectrometry using the minimum noise fraction transform. *BMC Research Notes* 2012 **5**:419.